DOCUMENT RESUME

ED 199 291 TM 810 202

AUTHOR Green, Kathy; Sax, Gilbert
TITLE Test Reliability by Ability Level of Examinees.
PUB DATE Apr 81
NOTE 12p.; Paper presented at the Annual Conference of the National Council on Measurement in Education (Los Angeles, CA, April 11-17, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Ability; *Achievement Tests; French; Higher Education; *Multiple Choice Tests; Teacher Made Tests; Test Construction; *Test Reliability

ABSTRACT

Achievement test reliability as a function of ability was determined for multiple sections of a large university French class (n=193). A 5-option multiple-choice examination was constructed, least attractive distractors were eliminated based on the instructor's judgment, and the resulting three forms of the examination (i.e. 3-, 4-, or 5-choice question form) were randomly assigned to quiz sections with similar mean cumulative grade point averages. Students were later grouped into high (3.6-4.0), average (3.1-3.5), and low (0-3.0) ability levels based on their final course grades in French where B=3.0 and A=4.0. A Kuder-Richardson 20 reliability coefficient was computed for each test form for each ability group and adjusted by the Spearman-Brown formula. Differences among reliabilities for the three forms were: (1) significant at alpha=.05 for the low ability group; (2) not significant for the high ability group; and (3) significant at alpha=.10 for the average ability group. The ability groups were combined and differences among reliabilities for the three forms were significant at alpha=.05. The optimal number of alternatives for all ability groups combined was four. (Author/RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED199291

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TEST RELIABILITY BY ABILITY LEVEL OF EXAMINEES

Kathy Green

and

Gilbert Sax

University of Washington

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

K. Green + G. Sax

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 810 202

Presented at the National Council on Measurement in Education, Annual Conference, Los Angeles, April 1981.

ABSTRACT

Achievement test reliability as a function of ability was determined for multiple sections of a large University of Washington French class. Previous empirical and theoretical papers suggested that reliabilities of tests with 3-option items were as high or higher than tests with 2-, 4-, or 5-options. Lord (1977), however, has argued that decreasing the number of options resulted in a more efficient test for high-level examinees but a less efficient test for low level examinees. Results of this study did not support this argument in a classroom situation. An explanation for the discrepancy is presented.

A number of studies have examined the effects on test reliability of the number of alternatives presented on multiple-choice items (Ebel, 1969; Grier, 1975; Lord, 1944, 1977). Several theoretical formulations have suggested that for integer values the 3-choice item allows maximum test reliability (Tversky, 1964; Grier, 1975) with 2-choice items next best (Grier, 1975). A model assuming knowledge or random guessing was used to algebraically derive the reliability of scores on a test composed of n equivalent A-choice items. This approach (Lord, 1977) suggested that 3-choice items are optimum in maximizing test reliability when difficult level p equals .5 and item intercorrelations r equal .2 or .3. Williams and Ebel (1957), Costin (1970, 1972), and Straton and Catts (1980) empirically found tests composed of 3-option items to be more reliable than 2-, 4-, and 5-option item tests. Lord (1977), however, using an item characteristic curve model with data from the College Board Scholastic Aptitude Test, found fewer options per item to be more efficient for high ability level examinees but to be less efficient for low ability level examinees when the total number of alternatives was held constant. Weber (1978) examined the effects of number of choices per item on reliabilities of classroom tests. She concluded that more choices per item yield higher test reliabilities for low achievers when time and test length are fixed. She compared only 3- vs. 4- and 3- vs. 5-choice tests with small sample sizes (N's=13-28) and short tests (19-20 items) in repeated administration of the tests to the same group. The present study compared the reliabilities of 3-, 4-, and 5-choice tests for low, average, and high ability level examinees. It examined whether the result suggested by Lord would be obtained under typical classroom testing conditions with use of a quasi-mastery exam rather than simulation of expected scores derived from the SAT as did Lord. It also provided an extension and improvement of Weber's design.

Weber used a repeated measures design, administering two versions of a test to the same group with a time lag between administrations. This design allows for confounding of results that are due to memory of item responses or to learning between test administrations. The present study employed independent groups.

Consistent with Lord (1977) and Weber (1978), the hypotheses for this study were:

1. Internal consistency reliability coefficients decrease significantly as number of options decreases for low-ability level examinees.
2. Reliabilities increase significantly as number of options decreases for high-ability level examinees.
3. No significant differences exist between reliabilities as number of options decreases for average-ability examinees.

METHOD

Participants in this study were 193 students in nine quiz sections of a beginning French class at the University of Washington. A 5-option multiple-choice examination was constructed. Distracters were then systematically eliminated from each question to form the 4- and 3-choice questions. Elimination of distracters was based on the instructor's judgment about the least attractive alternatives. The three forms of the examination were then randomly assigned to quiz sections (but not to individual students). Differences in mean cumulative grade point average (obtained from official academic records) among sections were assessed with a one-way analysis of variance. No significant main effect, however, was found for quiz section. All students within a given quiz section received the same test during the eighth week of instruction. Students were given 40 minutes to complete the 40-item tests. Since all students finished within this time,

speed was not considered to be a factor affecting performance. Students were later grouped into high, average, and low ability levels based on final course grades calculated independently of the results of the experimental exam. Grade point average cut-off points were chosen to provide approximately equal numbers of students in each ability group. The cut-off points were: high (3.6-4.0), average (3.1-3.5), and low (0-3.0).

RESULTS

Table 1 presents the item mean, test mean, and standard deviation for each test by ability group. A KR-20 reliability coefficient was computed for each test form for each ability group. To equate total number of items which could be given in the time used for a 5-option test, these reliabilities were then adjusted by the Spearman-Brown formula. This assumes that total testing time is proportional to the total number of alternatives, an assumption which is unlikely to be true for most item types but which is treated here as given. Adjusted and unadjusted reliability coefficients, number of items with non-zero variance, and sample sizes are presented in Table 2 for each ability level and for the combined sample.

(Tables 1 and 2 here)

Differences between reliability coefficients for groups and for test forms were tested with a statistic developed by Feldt (1969) and extended by Hakstian and Whalen (1976). The statistic (called "M") provides a test of the null hypothesis that reliability coefficients associated with k independent samples are equal and is based on the assumption that the scores on k parallel parts of a test conform to the assumptions of the two-factor random effects model of the analysis of variance: (1) a normally distributed population randomly sampled and (2) homogeneity of variance for the k parts of the test. Simulation studies suggest the test to be robust and slightly conservative (Hakstian & Whalen, 1976).

Differences among reliabilities for the low ability group for the 3-, 4-, and 5-choice tests were significant at α=.05 (M=10, df=2), but the trend of the reliabilities was clearly not the one hypothesized. In decreasing order of magnitude, the KR-20's favored the 4-choice test, the 3-choice test, and the 5-choice test. For the high ability group differences among reliabilities were not significant; for the average ability group differences were significant at α=.10 (M=5.94, df=2). Both of these last two results were contrary to hypotheses 2 and 3. The ability groups were combined and differences among reliabilities for the 3-, 4-, and 5-choice tests compared. These differences were significant at α=.05 (M=13.73, df=2). The optimal number of alternatives for all ability groups combined was four.

Differences in reliabilities among ability levels were also compared for each of the three tests. Differences were not significant ($p>.05$) for either the 3-, 4-, or 5-choice test.

DISCUSSION

Results suggest that a relatively easy teacher-made test may not conform to the theoretically reasonable predictions regarding test reliability of examinees of varying ability levels. Failure to support Lord's (1977) and Weber's (1978) results may derive from various factors: item means on the French tests deviated from the statistically optimal difficulty level of p=.5 (overall item mean in this study was $\bar{p}$=.78). The items were easier than those used in Lord's and in Weber's studies (median p = .5 and $\bar{p}$ = .65, respectively). The differences in item responses between ability groups may have been lessened since the item set was relatively easy.

Another difference between Lord's conditions and those in this study was the range of abilities available to categorize subjects as high, average, or low ability. Subjects in Lord's study had scaled scores on the 90-item verbal section

of the SAT ranging from 200 to 800. The range of abilities in the French class, as determined by final grade, was quite narrow: 72% of the class received at least a 3.0 for a final grade. Instead of presenting a contrast of low versus high ability, it is likely that this study contrasted moderately high with slightly higher ability levels on an easy test.

Another condition to consider is that tests were assigned randomly to quiz sections and not to individual students. Although quiz sections were not found to differ significantly in cumulative grade point averages, other systematic differences may have existed between sections.

In short, the conditions of this study differ from those idealistic conditions present in Lord's (1977) study. However, it is suggested that the conditions of this study--a fairly narrow range of abilities and a test with fairly easy items--are more representative of the typical classroom test than those in Lord's study which dealt with a more difficult test administered nationwide. It is interesting to note that the number of items with non-zero variance--the number of items a reliability coefficient is based upon--tended to decrease from the low to high ability groups. This would suggest that for easy tests, the reliability for high ability groups may tend to be depressed simply because of reduced variance among item responses in the high ability groups.

If achievement tests are designed to be relatively easy for a college class (e.g., $\bar{p} \geq .7$), it could be argued that items with fewer options would provide more efficient tests than items with more options. Ability range could probably be considered as homogeneous and narrow, abilities relative to the tested range being high. This argument would be supported by those empirical studies finding 3-option tests preferable to 4- and 5-option tests (e.g., Coston, 1970, 1972; Straton & Catts, 1980).

Table 1.

Item and Test Means and Standard Deviations by Ability Group*

Ability Group	3-choice $\bar{p}$	3-choice $\bar{X}$	3-choice SD	4-choice $\bar{p}$	4-choice $\bar{X}$	4-choice SD	5-choice $\bar{p}$	5-choice $\bar{X}$	5-choice SD
Low	.69	27.6	4.0	.69	27.8	6.5	.70	28.1	3.5
Average	.79	31.5	2.8	.81	32.6	4.3	.75	30.1	3.6
High	.85	33.9	3.0	.87	34.8	3.1	.85	34.0	3.0
Combined Sample	.77	30.9	4.2	.78	31.4	5.8	.77	30.9	4.3

*$\bar{p}$ was rounded to 2 digits; $\bar{X}$ and SD were rounded to 1 digit.

Table 2.

Adjusted and Unadjusted Kuder-Richardson 20 Reliability Coefficients by Ability Group and Number of Alternatives

Ability Group	3-choice				4-choice				5-choice			
	Unadj. KR-20	Adj. KR-20	Sample Size	Options ($s^2 \neq 0$)	Unadj. KR-20	Adj. KR-20	Sample Size	Options ($s^2 \neq 0$)	Unadj. KR-20	Adj. KR-20	Sample Size	Options ($s^2 \neq 0$)
Low	.55	.67	22	40	.84	.87	25	40	.43	.43	18	39
Average	.20	.30	22	38	.74	.78	18	34	.59	.59	9	30
High	.53	.65	20	35	.61	.66	20	33	.52	.52	18	31
Combined Sample	.66	.76	64	40	.85	.88	63	40	.68	.68	45	39

References

Costin, F. The optimal number of alternatives in multiple-choice achievement tests: some empirical evidence for a mathematical proof. Educational and Psychological Measurement, 1970, 30, 353-358.

Costin, F. Three-choice versus four-choice items: implications for reliability and validity of objective achievement tests. Educational and Psychological Measurement, 1972, 32, 1035-1038.

Ebel, R.L. Expected reliability as a function of choices per item. Educational and Psychological Measurement, 1969, 29, 565-570.

Feldt, L.S. A test of the hypotheses that Cronbach's Alpha or Kuder-Richardson Coefficient Twenty is the same for two tests. Psychometrika, 1969, 34, 363-373.

Grier, J.B. The number of alternatives for optimum test reliability. Journal of Educational Measurement, 1975, 12, 109-113.

Hakstian, A.R. & Whalen, T.E. A k-sample significance test for independent alpha coefficients. Psychometrika, 1976, 41, 219-231.

Lord, F.M. Reliability of multiple-choice tests as a function of choices per item. Journal of Educational Psychology, 1944, 35, 175-180.

Lord, F.M. Optimal number of choices per item--a comparison of four approaches. Journal of Educational Measurement, 1977, 14, 33-38.

Straton, R.G. & Catts, R.M. A comparison of two, three, and four-choice item tests given a fixed total number of choices. Educational and Psychological Measurement, 1980, 40, 357-365.

Tversky, A. On the optimal number of alternatives at a choice point. Journal of Mathematical Psychology, 1964, 1, 386-391.

Weber, M.B. The effect of choice format on internal consistency. Paper presented at the National Council on Measurement in Education Annual Meeting, Toronto, Canada, 1978.